Chapter 3
Compositing and Superposed Epoch Analysis

3.1 Introduction

Compositing, also sometimes called superposed epoch analysis when applied to time series, is one of the simplest analysis techniques, yet it can also be very powerful. It consists of sorting data into categories and comparing means for different categories. Although conceptually simple, compositing, like any other technique, must be applied logically, carefully and with proper concern for the possible appearance of spurious signals. Compositing is useful when you have many observations of some event and you are looking for responses to that event that are combined with noise from a lot of other influences. The idea is that if you average the data in some clever way in relation to the event, the event signal will remain and all other influences will tend to average out. Examples might include the climatic response to a volcanic eruption, the global weather response to el Niño, calculating the mean diurnal cycle of surface temperature in Dallas, Texas, or finding if precipitation responds to the phase of the moon. The last two of these relate to sorting out the true amplitude of cyclic responses. Often, compositing will reveal periodic phenomena with fixed phase that cannot be extracted from spectral analysis if the signal is small compared to the noise. Compositing makes no assumption of linearity, and it is good at separating small signals from noise, if your sample is large enough.

3.2 Steps in the Compositing Process

Setting up and executing a successful compositing study consists of several steps. If you think about these steps in the abstract, you will more easily see the ways to do it properly. It is easy to get excited with the specifics of a particular study and begin to lose your objectivity, especially if you hope to obtain a particular result. The steps are:

1. Select the basis for compositing and define the categories. The categories might be related to the phase of some cyclic phenomenon or forcing, or to time or distance from some event. Bases for compositing can range from the very commonplace, such as the hour of the day or the month of the year, to the relatively obscure, such as the passage of Earth through a reversal of the sun's magnetic field or the length of the solar cycle. It is highly desirable to have some believable hypothesis for why the event or cycle should affect the variables you are compositing. Otherwise you have greater risk finding a statistical coincidence with no physical meaning.
2. Compute the means and statistics for each category. One must be sure to be accurate and objective.
3. Organize and display the results. The results may be best shown in the form of tables, graphs or maps. It is important that whatever medium is used, a clear indication of the sense and significance of the results is achieved. This may mean adding confidence limits to the picture.

4. Validate the results. Validation of the results can be achieved in many ways. Statistical significance tests are only one of these. It is desirable to use as many of the following tests as possible, and any others that you can think of.
   - Use statistical significance tests. Using a model for the distribution of a variable about its mean, or nonparametric statistical tests, one can estimate the probability that the signal derived from the compositing exercise arose from chance. One should always do this type of testing, but it is not enough.
   - Subdivide the data set to show consistency. If you have enough data, you can divide the data set and see how well the derived relationships are maintained. This is especially useful if the statistical significance estimate is good, but the physical reason for the relationship is unclear. Monte Carlo techniques can be use here, *e.g.* divide the sample many ways randomly. How often does the result reproduce?
   - Show consistency in other ways.
     - Find some additional data and reproduce the results
     - Show that the results are consistent with space or time
     - Show that the results are consistent with a well-founded theory

3.3 Evaluating compositing studies

When evaluating a compositing study, ask yourself the following questions.

1. Do you have an *a priori* basis for expecting to find the relationships found in the study? *A priori* means beforehand: based in knowledge or hypothesis that was available to the investigator before the study was conducted. If there was not, then you might suspect that the relationship was found after trying a number of different things, in which case the probability that it occurred by chance is greatly increased, and *a priori* statistical tests should not be used. Every time you give your hypothesis another independent chance to succeed you need to multiply your certainty by the *a priori* probability that it is true (1.10).
2. What is the basis for compositing? Does it have a precise, unique, objective definition (*e.g.* time of day) or is it somewhat arbitrary? Does it have a distinct physical interpretation? Could another basis for compositing have been used just as well, or a better one defined?
3. How was the compositing performed? Can you easily visualize how the process might have been programmed for the computer? Could an opportunity for subjective judgment or subconscious bias have entered the procedure at some point?
4. How does the investigator argue that the results are statistically significant (*i.e.* that they would be reproducible in independent data sets)? List all the statistical arguments that are given and the physical or logical arguments as well. Can you think of alternative, perhaps simpler, explanations? Are there reasons to suspect that the method itself produced a signal that is not really in the data? Does the author have a justification for using *a priori* statistical tests?
5. Are you convinced of the validity of the results? Would you direct your research effort on the assumption that the results are correct? If not, what would it take to convince you?

3.4 Example: Daily Precipitation and Temperature

We download daily station data from the U.S. Climate Data Center. We choose the location of the Seattle SeaTac International Airport, which has records from January 1, 1948 to the present, about 69 years. We organize the data as a two-dimensional array by year from 1948 to 2016 and by day of the year from 1

(January 1) to 365 (December 31). We ignore the extra day in leap years. We compute the mean and standard deviation for each day of the year, using the sample of 69 years. We then plot the mean as a function of day of year.

Fig. 3.1 shows the raw mean of the maximum daily temperature and precipitation for the sample as a thin black line. Since the data are still noisy even with a sample of 69 years, we also show two different locally weighted scatterplot smoothings (LOWESS) of the data, with the red line showing a stronger smoothing. Also shown are dashed lines, which are the less smoothed data, plus or minus the smoothed standard deviation for each day. We see that the standard deviation of temperature is similar in winter and summer. Later we will show that it is actually a little bigger in summer.

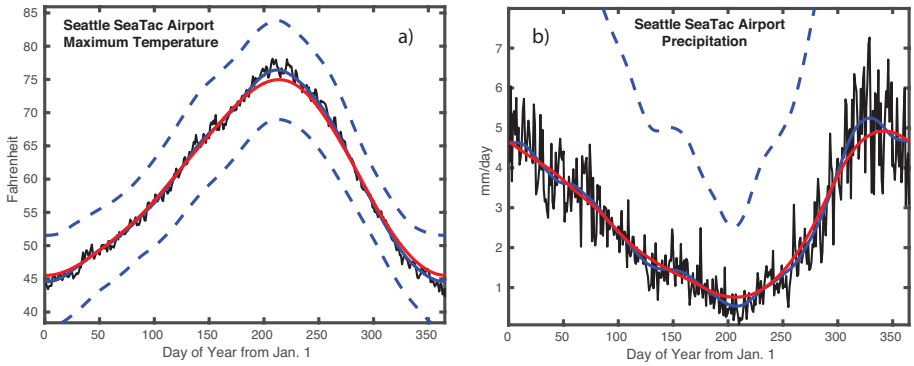

Figure 3.1 Composited a) daily maximum temperature and b) precipitation for 69 years of data from SeaTac Airport near Seattle, Washington. Blue and Red lines are two different smoothings of the daily composites. Dashed lines indicate plus or minus one smoothed standard deviation.

The annual cycle composite for daily precipitation shows that it rains more in the winter, and almost not at all for a brief period near July 31st (day 212). In this case we see that the standard deviation is larger than the mean, and that subtracting the standard deviation gives a non-physical negative value, which is not shown. This is because precipitation is not a normally distributed variable, since it has a minimum value of zero, which is also the most commonly occurring value. The skewness for the SeaTac maximum temperature data is 0.4 in July and -0.5 in January, both very close to the value of 0.0 for a normally distributed variable, while the kurtosis is 2.8 in July and 3.5 in January, also close the the normal distribution value 3.0. The precipitation data on the other hand have skewness of 5.2 in July and 3.0 in January, and kurtosis of 35.2 in July and 15.6 in January.

Fig. 3.2 shows the probability density functions for maximum daily temperature and daily precipitation for July and January obtained from the data by the kernel method. Temperature looks fairly Gaussian like a normally distributed variable, while the precipitation pdf peaks near zero and is better fit with a gamma distribution. The July maximum temperature pdf is wider and less peaked than the December one, indicating warmer, but more variable temperature in the summer. Because precipitation is less frequent in the summer, its pdf is even more strongly peaked at zero in July than January.

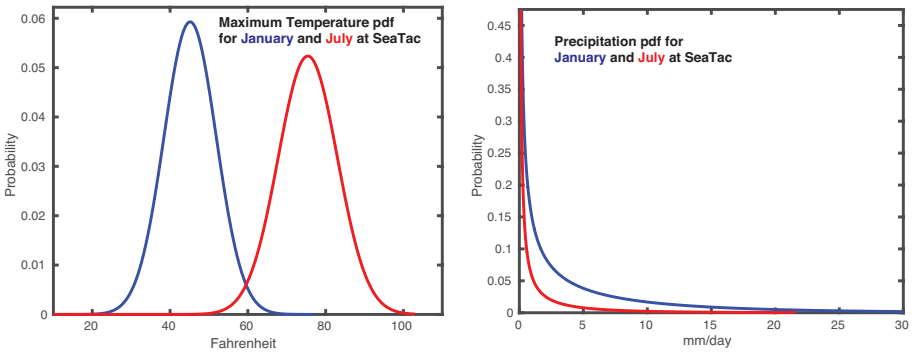

Figure 3.2 Smoothed Probability Density Functions for daily maximum temperature and precipitation for SeaTac Airport during July and January. Note that the precipitation pdf is unbounded at zero, but is cut off at 0.45 and at 30 mm/day so that the differences in the pdfs for precipitation can be better seen.